Sentimentogram: A Personalized Subtitle System Using Multi-modal Emotion Recognition and Visualization

Anonymous submission

Abstract

This paper presents Sentimentogram, an innovative emotion-aware subtitle generation system that reimagines subtitles as adaptive, emotionally expressive interfaces rather than static text overlays. Leveraging a multimodal Speech Emotion Recognition architecture, our system fuses acoustic representations from Wav2Vec2 with linguistic features extracted via BERT, employing self-attention and cross-attention mechanisms to capture the nuanced interplay of tone and semantics. Detected emotions are projected into the Valence-Arousal-Dominance (VAD) space and rendered visually through dynamic alterations in subtitle color, font size, and style. Personalization is achieved through a multi-layered framework that adjusts subtitles according to user age, cultural background, and playback conditions, optimizing both accessibility and emotional resonance. Our model achieves competitive performance on the IEMOCAP benchmark, validating its effectiveness against leading SER approaches. By integrating affective intelligence with cultural adaptability, Sentimentogram offers a new paradigm for emotionally immersive and inclusive media experiences, particularly vital for users in acoustically constrained or hearing-impaired contexts.

1. Introduction

The past decade has witnessed remarkable advances in Speech-to-Text (STT) technology, enabling machines to transcribe human speech with near-perfect precision. Yet, beneath this technical triumph lies a fundamental omission: the absence of emotion, a core facet of human communication. While contemporary STT systems adeptly capture what is said, they remain largely incapable of discerning how it is said—missing the subtle intonations of joy, sorrow, tension, or urgency that imbue spoken language with emotional richness. The result is often a semantically accurate yet affectively impoverished transcription.

This shortcoming extends poignantly to media subtitles. Stripped down to static strings of text, traditional subtitles frequently fail to convey a speaker's emotional undertones or the atmospheric mood of a scene. Such flattening of expressive content creates a perceptual disconnect, limiting the viewer's emotional engagement and diminishing narrative immersion. To address this gap, we propose a personalized, emotion-aware subtitle generation system that fuses high-fidelity STT with advanced emotion recognition. Accurate emotional inference depends not only on linguistic information but also on paralinguistic cues—namely, prosodic and acoustic features—which necessitate a multimodal framework. Models constrained to a single modality are ill-equipped to parse the nuanced emotional dynamics of natural conversation, where meaning emerges from the interplay between words and vocal tone. Moreover, conventional subtitle systems are inherently static and context-insensitive, offering no accommodation for viewers' demographic profiles, situational contexts, or individual preferences.

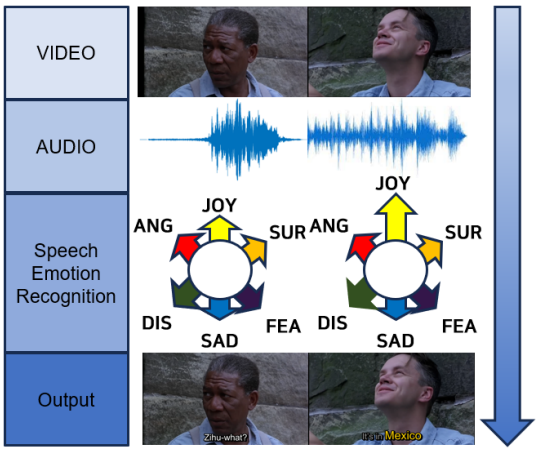

Figure 1: Overall pipeline for emotion-aware subtitle generation. Given input video frames and corresponding audio, a speech emotion recognition module predicts one of six basic emotions (Anger, Disgust, Fear, Sad, Surprise, Joy), which is then mapped to a predefined color/style scheme. The right-hand example shows the same subtitle rendered with emotion-based coloring "Mexico." in yellow for Joy.

To overcome these limitations, our study explores the cross-modal relationship between emotion and color, a topic long examined in cognitive science and cross-cultural psychology. A large-scale international study (Jonauskaite et al., 2020) demonstrates that emotion–color mappings such as red→anger and black→sadness reflect universal cognitive structures across cultures, while also highlighting cultural differences—for example, white signifies purity in Western cultures but mourning in many East Asian traditions.

Building on these insights, we propose Sentimentogram, an emotion-aware subtitle generation system featuring:

1. A high-precision STT engine (e.g., Whisper) for accurate speech transcription

2. A multimodal neural framework that fuses acoustic and textual embeddings via cross-attention

3. A dynamic typographic layer that adjusts font size according to emotional intensity

4. A personalization engine that applies emotion–color mappings and style adaptations tailored to user characteristics (age, culture, playback conditions)

To our knowledge, this is the first comprehensive effort to integrate affective intelligence and contextual sensitivity into subtitle systems. By simultaneously reflecting universal emotional cognition and cultural particularities, Sentimentogram redefines the viewing experience, making it more immersive and personalized. It also holds great promise for conveying emotional nuance when auditory access is limited, thereby enhancing accessibility and inclusivity in digital media.

2. Related Works

2.1 Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) has long prioritized improving classification accuracy and model robustness. For instance, emoDARTS automates architecture search via DARTS to jointly optimize CNNs and sequential networks, achieving state-of-the-art results on IEMOCAP (Rajapakshe et al. 2024). Similarly, integrating self-supervised pretraining with speaker normalization has been shown to reduce speaker-dependent variability and enhance generalization across diverse real-world speech data (Gat et al. 2022).

More recently, contrastive feature reconstruction and aggregation has been applied to multi-modal, multi-label emotion recognition. The CARAT framework uses modality- and label-specific contrastive losses together with shuffled aggregation to capture fine-grained co-occurrence patterns, outperforming prior fusion methods on CMU-MOSEI and M³ED benchmarks (Peng et al. 2024). In parallel, emotion2vec introduced at ACL 2024 Findings demonstrates that self-supervised pretraining on large unlabeled speech corpora, followed by linear probing, significantly outperforms Wav2Vec2 on IEMOCAP and multilingual SER tasks (Ma et al. 2024).

Despite these advances in classification, fusion, and pretraining, few studies have addressed how to dynamically integrate recognized emotion signals into media content—particularly real-time subtitle rendering. To bridge this gap, we propose the 'Sentimentogram' framework.

2.2 Models for Subtitles

The most fundamental technology for automatic subtitle generation is accurate Speech-to-Text (STT). A representative model in this field is OpenAI's Whisper, which was trained using weak supervision on a large and diverse audio dataset, resulting in exceptional robustness to background noise, various accents, and technical jargon (Radford et al. 2023). Thanks to these characteristics, Whisper has established itself as a powerful foundational technology for extracting reliable text from real-world video content. Building on such high-precision STT technology, research is expanding beyond simple transcription toward enhancing the expressive power of subtitles. For example, one study proposed a method using Generative Adversarial Networks to generate contextually appropriate and stylistically diverse video subtitles (Shen 2023; Toshpulatov et al. 2025).

However, while these approaches contribute to enhancing the linguistic richness and naturalness of subtitles, they have not yet extended to personalizing subtitles by reflecting non-linguistic and personal factors such as the speaker's underlying 'emotion' or the viewer's specific 'context' (e.g., age, viewing preferences, playback speed).

Therefore, this study aims to fill the gap identified in the prior research. To this end, we intend to propose a new, integrated approach that fuses SER technology with personalized subtitle generation technology to consider both the speaker's emotion and the viewer's context.

3. Methodology

3.1 VAD-Based Emotion-Color Mapping

The role of colors in emotional perception and communication is significant, as they provide a visually intuitive way to express and interpret emotions (Valdez and Mehrabian 1994). The connection between colors and emotions is shaped by psychological principles and cultural contexts, making it an important area of study across multiple disciplines (Russell 1977).

A three-dimensional framework for quantifying emotions—Pleasure–Arousal–Dominance (PAD)—was first introduced in 1996 to systematically describe affective states (Mehrabian 1996). The pleasure–displeasure axis differentiates between positive and negative emotions, while the arousal–non-arousal dimension represents activation intensity, ranging from high-energy excitement to a relaxed state . A third axis captures the degree of control or autonomy experienced by an individual. Subsequent work demonstrated that these three independent dimensions suffice to characterize a broad spectrum of emotions (Russell 1977). Empirical research further linked these emotional dimensions to the physical properties of color—hue, saturation, and brightness.

Building on this foundation, the Valence, Arousal, Dominance (VAD) model—standard in affective computing—represents emotions as continuous numerical vectors, enabling nuanced distinctions (e.g., "bittersweet") that categorical labels cannot capture. Its vector form also makes it well suited for quantitative tasks such as feature input in machine-learning models or similarity computations between emotional states. Formally, the VAD model defines an emotion vector E as:

$$\mathbf{E} = (V, A, D) = f(X; W) \tag{1}$$

where:
V:The Valence value,
representing the positive-negative dimension
A:The Arousal value,
representing the activation–relaxation dimension,
D:The Dominance value

Emotion	Valence(V)	Arousal(A)	Dominance(D)
Joy	0.960	0.648	0.588
Anger	–0.666	0.730	0.314
Sadness	–0.896	–0.424	–0.672
Fear	–0.854	0.680	–0.414
Surprise	0.750	0.750	0.124
Disgust	–0.896	0.550	–0.366
Neutral	–0.062	–0.632	–0.286

Table 1: The 7 core emotions based on Ekman's basic emotions and their corresponding VAD values extracted from a large-scale lexicon.

representing the control-submission dimension
X:The feature vector of the input speech signal
W:Denotes model parameters

This study adopts an integrated approach, using a basic classification system of seven emotions (Ekman's six basic emotions add a 'Neutral' state)(Ekman 1992) and then mapping them into the VAD space. This allows us to leverage the advantages of both intuitive emotion classification and nuanced visual representation.

These quantified VAD values are used directly to determine the visual properties of the subtitles. Specifically, Valence is mapped to Hue, Arousal is mapped to Saturation, and Dominance is mapped to Lightness and font weight, allowing each emotion to be systematically translated into a unique color and style. It is noteworthy that the results of this systematic mapping show a high degree of consistency with the universal color-emotion associations found in the large-scale survey by (Jonauskaite et al. 2020). This suggests that the methodology of this study, despite using a quantitative model, produces results similar to general human perception, thereby securing its validity.

3.2 Speech Emotion Recognition (SER)

The Speech Emotion Recognition (SER) in this study adopts a multi-modal approach that leverages both acoustic and textual features inherent in the speech signal to enhance the accuracy of the recognition process. This is to reduce the ambiguity that can arise when relying on a single modality and to better capture the multifaceted nature of human emotional expression. The overall architecture consists of a Joint Audio-Text Encoder, which extracts and fuses acoustic and textual features, and an Emotion Classifier, which determines the final emotion based on the fused features.

In the acoustic feature extraction stage, a pre-trained Wav2Vec2 model is used as the encoder. This model, through self-supervised learning, directly extracts high-dimensional contextual acoustic feature sequences, $\mathbf{F}_{\text{Audio}} \in \mathbb{R}^{B\times N\times D}$, from the raw audio waveform. It is then fine-tuned to learn acoustic representations specialized for speech emotion classification. Concurrently, in the textual feature extraction stage, the speech signal is first converted into text subtitles via OpenAI Whisper (Radford et al. 2023). The generated text is then input into a powerful pre-trained language model, BERT, to extract deep semantic and contextual features, resulting in a feature sequence $\mathbf{F}_{\text{Text}} \in \mathbb{R}^{B\times N\times D}$ that matches the dimension of the acoustic features(Devlin et al. 2018). Next, the acoustic feature sequence $\mathbf{F}_{\text{Audio}}$ and the textual feature sequence $\mathbf{F}_{\text{Text}}$ are concatenated along the sequence dimension to form a single unified sequence, X:

$$\mathbf{X} = \left[\mathbf{F}_{\text{Audio}};\, \mathbf{F}_{\text{Text}}\right] \in \mathbb{R}^{B\times(N_s+N_t)\times D}. \tag{2}$$

This unified sequence X is then input into a Self-Attention layer for Joint Encoding. This process allows the acoustic and textual features to mutually reference each other's context to learn global relationships. The final joint representation $\mathbf{H}$ is obtained by applying a residual connection and layer normalization to the output of the Self-Attention:

$$\begin{aligned}\mathbf{H}' &= \text{SelfAttention}\big(\mathbf{X_q}, \mathbf{X_v}, \mathbf{X_k}\big),\\ \mathbf{H} &= \text{Layer}\big(\mathbf{X} + \mathbf{H}'\big).\end{aligned} \tag{3}$$

Afterward, the joint representation $\mathbf{H}$ is split back into its acoustic part, $\mathbf{H}_A \in \mathbb{R}^{B\times N_s\times D}$, and its textual part, $\mathbf{H}_T \in \mathbb{R}^{B\times N_t\times D}$. These two representations are then deeply fused through multi-head cross-attention. First, an audio-to-text attention learns which parts of the text the audio should focus on, and second, a text-to-audio attention learns which parts of the audio the text should focus on.

$$\begin{aligned}\mathbf{H} &= [\mathbf{H}_A;\, \mathbf{H}_T],\\ \mathbf{H}'_{A\to T} &= \text{MHCA}\big(\mathbf{Q}=\mathbf{H}_A,\ \mathbf{K}=\mathbf{H}_T,\ \mathbf{V}=\mathbf{H}_T\big),\\ \mathbf{H}'_{T\to A} &= \text{MHCA}\big(\mathbf{Q}=\mathbf{H}_T,\ \mathbf{K}=\mathbf{H}_A,\ \mathbf{V}=\mathbf{H}_A\big).\end{aligned} \tag{4}$$

Finally, the two cross-attention outputs are combined again to create the final multimodal representation, $\mathbf{Y}$. The role of this step is to synthesize the refined information from both directions to create the final feature to be used by the classifier. The resulting fused representation $\mathbf{Y}$ is then fed into an MLP classifier, which maps it to one of the seven emotion classes, and the entire model is trained via a cross-entropy loss function.

$$\mathbf{Y} = \left[\mathbf{H}'_{A\to T};\, \mathbf{H}'_{T\to A}\right] \tag{5}$$

The final fused representation $\mathbf{Y}$ obtained from the previous stage, is first summarized into a single vector that represents the entire utterance, before being input to the Classification Head.

$$\bar{h} = \frac{1}{N_s+N_t}\sum_{i=1}^{N_s+N_t} Y_i \tag{6}$$

The pooled vector $\bar{h}$ is passed through a linear layer to compute the final logits $\mathbf{z} \in \mathbb{R}^{B\times C}$ where $W_c \in \mathbb{R}^{D\times C}$ is the weight matrix, $\mathbf{b}_c \in \mathbb{R}^C$ is the bias vector, and C denotes the number of emotion classes.

$$\mathbf{z} = \bar{h}\, W_c + \mathbf{b}_c \tag{7}$$

The calculated logits $\mathbf{z}$ are passed through a Softmax function to obtain the final probability distribution $\mathbf{P} = \text{Softmax}(\mathbf{z})$ for each emotion class. The final loss for model training is calculated using the standard cross-entropy loss, defined as $\mathcal{L} = -\sum_{c=1}^{C} y_c \log P_c$, which measures the difference between the predicted distribution $\mathbf{P}$ and the true

labels $\mathbf{y}$.

$$P_{b,c} = \frac{\exp\big(z_{b,c}\big)}{\sum_{k=1}^{C} \exp\big(z_{b,k}\big)} \tag{8}$$

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c \, \log P_c \tag{9}$$

3.3 Personalized Subtitle

The core objective of this system is to move beyond the conventional static approach of providing identical subtitles to all users, and instead offer genuinely personalized subtitles that dynamically react to user characteristics and the viewing environment. To achieve this, 'Sentimentogram' implements its personalization strategy along three main axes: Age, Culture, and the Viewing Environment.

User age is a key demographic variable that significantly influences linguistic preferences and content receptivity. To provide subtitles optimized for different age groups, this system applies a Synonym Mapping technique using a semantic network such as WordNet from the NLTK (Natural Language Toolkit). The core of this technology is to quantitatively measure the semantic similarity between words, for which it uses Wu-Palmer Similarity (Wu and Palmer 1994). This method calculates the relative depth of two words within the WordNet hierarchy by utilizing the depth of their Lowest Common Ancestor (LCA) (Liu et al. 2025), and is defined as follows:

$$\mathrm{sim}_{\mathrm{wup}}(w_1, w_2) = \frac{2\,\mathrm{depth}\big(\mathrm{LCA}(w_1, w_2)\big)}{\mathrm{depth}(w_1) + \mathrm{depth}(w_2)}\,. \tag{10}$$

The system classifies users into three groups—children, teenagers/young adults, and adults—and provides the following differentiated subtitles based on this technology.

For the Children group, 'Censored Subtitles' are provided for a safe viewing experience, using a 'Censorship Module' based on the NLTK library's profanity list to mask inappropriate words.

For the Teenagers and Young Adults group, 'Slang Subtitles' are provided to enhance relatability. This feature utilizes the synonym mapping technology described earlier to transform a standard word like 'excellent' into a semantically similar colloquial expression such as 'awesome,' based on its Wu-Palmer Similarity score.

For the Adults group, 'Standard Subtitles' are provided by default, without any filtering or transformation, to ensure the clear delivery of information.

While the system's emotion visualization is based on universal tendencies, true personalization is achieved by respecting a user's cultural background. The theory of psychological construction of emotion posits that emotions are 'constructed' via cultural knowledge, implying that symbolic systems like color must also vary by culture. Ignoring these differences leads to what this study defines as 'Affective Incongruity': a conflict between the system's intended visual meaning and the user's cultural interpretation. As demonstrated by (Jonauskaite et al. 2020), while universal patterns exist, significant cultural differences are also present; for example, 'sadness' is associated with black in the West but with white in China. This incongruity can cause psychological 'design dissonance' (Festinger 1957; Reiss 2013; Toshpulatov et al. 2024) and act as an 'affective mis-translation' that undermines the system's communicative purpose.

To solve this problem, this system implements an intelligent personalization architecture based on the principles of culturally adaptive interfaces from HCI (Reeves and Nass 1996). This architecture consists of a 'Cultural User Model,' a 'Cultural Knowledge Graph' that stores the rules (Pussadeniya, Nakisa, and Rastgoo 2024), and a 'Dynamic Rendering Engine' that renders subtitles according to the queried rules. For instance, if a Chinese user is detected and 'sadness' is recognized, the system references the 'Chinese' profile and renders the subtitle in 'light gray' instead of the default blue or black.

Finally, since the Playback Speed critically impacts readability, the system dynamically adjusts the subtitle format to account for modern viewing habits where high-speed playback is common. To achieve this, the system adaptively alters the information density of the subtitles. The core mechanism is to flexibly adjust the weight between full-text subtitles and emoji-based summary subtitles. The emoji weight (W_{emoji}) corresponding to the playback speed (s) is determined by the following normalization function

$$W_{\mathrm{emoji}}(s) = \max\Big(0,\ \min\Big(1,\ \frac{s_{\min} - s_{\max}}{s_{\max} - s_{\min}}\Big)\Big) \tag{11}$$

In a high-speed environment (s ¿ 1.5x), the (W_{emoji}) value approaches 1. In this case, the system provides a visual cue by prepending a representative emoji, which signifies the emotion recognized by the SER module, to the beginning of the original subtitle text. For example, the sentence "I'm so happy today!" is transformed into " I'm so happy 😊 today!". This allows the user to quickly grasp the emotional nuance from the emoji in advance, before reading the full sentence, enabling a deeper understanding of the content.

Conversely, in a normal or low-speed environment, the W_{emoji} value approaches 0, and the full text of the subtitle is displayed. In this situation, the system optimizes for readability by maintaining a target CPS (Characters Per Second) of approximately 15 to ensure a comfortable reading pace for the user. If the initial subtitle generated via STT (Speech-to-Text) exceeds a CPS of 15 and is thus too fast, the system may auto-split it into two shorter subtitles or adjust the timing with adjacent subtitles to secure enough display time. If the CPS is too low, causing the subtitle to linger unnecessarily, the system may display the next subtitle slightly earlier to improve the overall flow and pacing. Through this CPS-based dynamic time adjustment, the user can consistently read the subtitles at a comfortable speed under any condition, allowing them to fully understand the content and appreciate the emotional nuances.

Through these approaches, 'Sentimentogram' implements a flexible and scalable personalization that respects both the universality and the cultural specificity of emotional

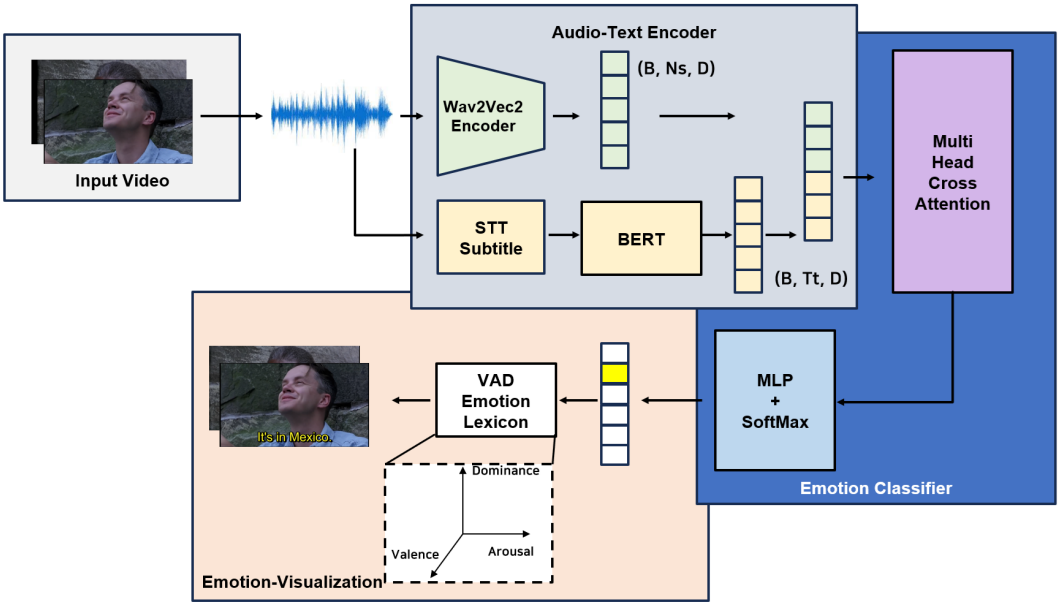

Figure 2: Overview of the proposed Sentimentogram architecture. Acoustic features from Wav2Vec2 and textual features from BERT are fused via self- and cross-attention modules to generate emotion-aware subtitle visualizations.

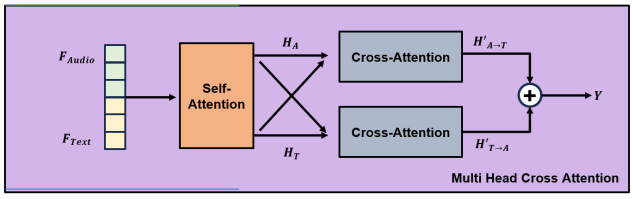

Figure 3: Detailed view of the bi-directional cross-attention module, showing how audio and text features attend to each other.

expression. This aims to provide a media experience that is not only functionally accurate but also culturally inclusive, ethical, and ultimately creates a deeper emotional resonance for diverse users worldwide.

4. Experiments

This chapter quantitatively evaluates the performance of the multi-modal emotion recognition model proposed in Chapter 3. To situate our model's performance within the current research landscape, we use standard benchmarks in the field of affective computing.

4.1 Datasets

IEMOCAP (Interactive Emotional Dyadic Motion Capture): The primary dataset for training and evaluating our proposed model is IEMOCAP, a multi-modal database of dyadic conversations featuring approximately 12 hours of data from ten professional actors (Busso et al. 2008). It includes both scripted and improvised interactions. For this study, we utilized the audio and transcribed text to train and evaluate our model on the six-class emotion task (happiness, sadness, neutral, anger, excitement, and frustration).

Condensed Movies Dataset: For domain adaptation, we employed a second dataset. The Condensed Movies dataset consists of clips extracted from various film genres and was used to fine-tune the model further, enhancing its robustness against complex acoustic environments that include background music and sound effects (Bain et al. 2020).

4.2 Evaluation Metrics

The SER model's performance was assessed using a suite of standard classification metrics, including Accuracy, Precision, Recall, and F1-Score. We specifically focus on Weighted Accuracy (WA), to understand the overall performance, and Weighted F1-Score (WF1), which provides a more balanced measure by accounting for the number of samples in each class.

4.3 Result and Analysis

To demonstrate the superiority of the proposed multi-modal model, its performance was compared against several state-of-the-art (SOTA) studies on the IEMOCAP six-class emotion task. Recent SOTA research has advanced around three key architectural paradigms: Graph Neural Networks (GNNs) for modeling conversational structure, novel approaches leveraging the inferential power of Large Language Models (LLMs) by converting speech features into natural language prompts, and Transformer-based multi-modal fusion techniques that fuse speech and text, as in the present study.

[Table 2] summarizes the performance of major models based on these recent methodologies. As for the individual approaches, GASMER (2025) recorded a WF1 of 71.20% through adaptive graph structure learning. SpeechCueLLM (2024) achieved a WF1 of 72.60% by converting the

Model	Year	Key Architecture	WA (%)	WF1 (%)
GASMER	2025	GNN, Adaptive Structure Learning	–	71.20
SpeechCueLLM	2024	LLM-Prompt (LLaMA-2)	–	72.60
GraphSmile	2024	GNN, RoBERTa	–	72.77
SSEAN	2025	Dual Recurrent Networks, Commonsense Knowledge	–	73.94
SDT	2023	Transformer, Self-Distillation	74.08	–
Ours	–	Self & Cross-Attention Fusion	**76.03**	**76.01**

Table 2: Performance comparison on the IEMOCAP benchmark of weighted accuracy (WA) and weighted F1-score (WF1) for several state-of-the-art SER models. Models are listed with their publication year and key architectural components. Our proposed Cross-Attention Fusion approach achieves 76.0 % WA and 76.0 % WF1, outperforming prior methods by 2–4 points in weighted F1.

Model Configuration	**Accuracy (%)**
w/o Audio (Text-only)	67.0
w/o Text (Audio-only)	69.0
Simple Fusion (Concat)	69.0
Proposed Model (Ours)	76.0

Table 3: Ablation study on the IEMOCAP benchmark evaluating the impact of each modality and fusion strategy.

prosodic features of speech into a natural language description and inputting it as a prompt to an LLM along with the dialogue content. GraphSmile (2024) modeled the dynamic relationships between speakers using a GNN to achieve a WF1 of 72.77%. SSEAN (2025) recorded a WF1 of 73.94% with a unique approach that separately models scene-level and speaker-level emotions and integrates external commonsense knowledge. Finally, SDT (2023) achieved a high Weighted Accuracy (WA) of 74.08% by applying a self-distillation technique to its Transformer architecture to enhance the representation of each modality. In comparison, the model proposed in this study, which deeply fuses acoustic and textual features through Self-Attention and Cross-Attention, achieved a Weighted Accuracy (WA) of 76.0% and a Weighted F1-Score (WF1) of 76.0%, demonstrating highly competitive performance against the existing SOTA models presented in the table.

4.4 Ablation Study

To see how each component contributes, we ran an ablation study that starts from Text-only and Audio-only baselines and then adds fusion steps. As Table 3 shows, Text-only reaches 67.0% accuracy and Audio-only 69.0%. Naively concatenating the two (Simple Fusion) also hits 69.0%, matching the audio-only result and demonstrating that mere concatenation can't exploit cross-modal interactions. Once we introduce both self- and cross-attention, accuracy jumps to 76.0%, a 7-point gain. That jump makes it clear that our attention-based fusion is key to capturing the interplay between acoustic and textual cues.

4.5 Personalized Subtitle Evaluation

To evaluate the practical effects of the proposed personalized subtitle system ('Sentimentogram') on user experience, a user study was conducted. Participants watched the same

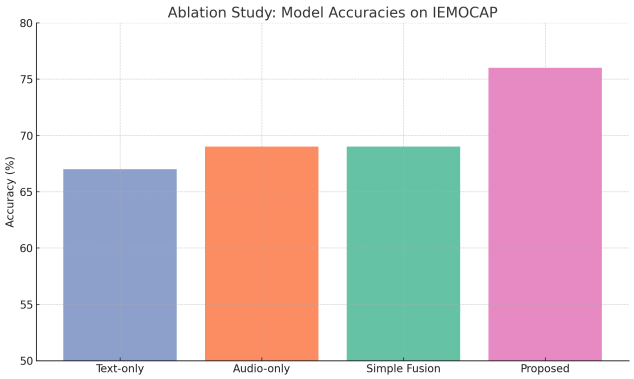

Figure 4: Ablation Study: Model Accuracies on IEMOCAP. Accuracy (%) for Text-only, Audio-only, Simple Fusion and Proposed models, plotted on a 50–80 range to highlight performance differences.

video clip with both regular subtitles and subtitles generated by the proposed system. Afterward, they rated the subtitles on multiple metrics, including Clarity, Emotion Reflection, and Satisfaction.

[Table 4] summarizes the mean user evaluation scores for each metric. The analysis revealed that the proposed system scored higher than regular subtitles across all metrics, with the most significant performance improvements seen in areas directly related to its core features.

The most noteworthy results were the substantial increases in the 'Emotion Reflection' score from 2.4 to 3.8 (+1.4) and the 'Color Accuracy' score from 2.8 to 4.6 (+1.8). This strongly suggests that the VAD-based emotion visualization framework designed in Chapter 3 effectively conveys the speaker's emotional state to the viewer.

This enhanced emotional delivery also had a positive impact on the overall user experience. User 'Engagement' rose significantly from 3.2 to 4.8 (+1.6), 'Visual Appeal' from 3.2 to 4.6 (+1.4), and 'overall Satisfaction' from 3.6 to 5.0 (+1.4), with all metrics showing a large increase.

Furthermore, scores also improved for basic metrics such as Clarity, Comprehension, and Readability. This indicates that the emotion visualization and personalization features provided a richer media experience without hindering the fundamental information-delivery function of the subtitles.

In conclusion, the user evaluation empirically demon-

Metric	Regular	Generated	(Gen–Reg)
Clarity	3.8	4.2	+0.4
Comprehension	3.4	4.2	+0.8
Effort	3.4	4.2	+0.8
Readability	3.8	4.6	+0.8
Emotion Reflection	2.4	3.8	+1.4
Color Accuracy	2.8	4.6	+1.8
Engagement	3.2	4.8	+1.6
Visual Appeal	3.2	4.6	+1.4
Satisfaction	3.6	5.0	+1.4
Recommendation	3.4	4.2	+0.8

Table 4: Mean user evaluation scores on a 1–5 scale comparing standard subtitles and subtitles generated by our model across multiple metrics. Subtitles generated by our model show superior performance in emotion reflection, color accuracy, and overall engagement, demonstrating the impact of emotion-aware visualization on subtitle quality.

strates that the proposed 'Sentimentogram' system effectively visualizes emotions and significantly enhances user immersion and satisfaction, successfully overcoming the limitations of conventional subtitle systems.

5. Conclusion and limitation

This study introduces Sentimentogram as a response to a foundational shortcoming in contemporary media consumption, namely, the emotional disconnect perpetuated by conventional, static subtitles. Moving beyond the limitations of rudimentary text transcription, which fails to capture the rich emotional nuances embedded in human speech and lacks any personalization mechanism, we present a novel paradigm. By integrating a highly sophisticated multimodal emotion recognition architecture with a multi-layered personalization framework, we construct a truly adaptive subtitle system tailored to both content and context.

At the core of our system lies a robust Speech Emotion Recognition (SER) model, which deeply fuses acoustic signals with linguistic context through Self-Attention and Cross-Attention mechanisms. This methodological design demonstrated competitive efficacy, achieving 76.0% accuracy on the IEMOCAP benchmark, positioning it favorably among state-of-the-art approaches and validating the architectural soundness of our solution. Importantly, while our system benefits from strong SER performance, its personalization logic remains fundamentally model-agnostic, reinforcing its versatility.

Building upon this robust emotional foundation, we engineered a dynamic personalization engine. This component adaptively modifies subtitles through a combination of age-specific language adjustments, culture-sensitive color mapping, and readability optimization according to playback speed, resulting in a nuanced and context-aware subtitle experience.

Ultimately, Sentimentogram reconceptualizes subtitles not as a static textual overlay, but as an intelligent and responsive interface, one that adapts fluidly to the emotional tenor of the content and the personal context of the viewer. By pioneering a new method for visualizing and tailoring emotional expression, this research proposes a blueprint for a future in which media consumption is more immersive, inclusive, and emotionally resonant. Furthermore, our system demonstrates clear utility in scenarios where auditory access is limited, such as environments with ambient noise or for individuals with hearing impairments, positioning it as a critical modality for equitable media accessibility.

Nevertheless, despite the successful validation of emotion-driven personalized subtitle generation, several limitations warrant discussion. First, the system's dependence on the Speech-to-Text (STT) component introduces potential fragility in the user experience. Our implementation adopts Whisper as the transcription backbone, which, despite its state-of-the-art performance, remains fallible. Errors in transcription can have two significant consequences: (1) misrecognition of emotionally salient words may impair the downstream emotion classifier's accuracy; and (2) more critically, transcription errors are directly manifested in the subtitle output, undermining the core function of accurate information delivery. This dual failure can disrupt user immersion and erode trust in the system. It is important to note, however, that our proposed framework is not intrinsically tied to any specific STT engine, and can be decoupled or enhanced as more robust transcription models emerge.

Second, the scope of cultural personalization remains limited in its current instantiation. The implemented Cultural Knowledge Graph relies on well-documented emotional-color associations from approximately 30 nations. While this serves as a valuable proof-of-concept, a truly global system necessitates far more comprehensive data collection and systematic expansion of cultural mappings to capture the full spectrum of cross-cultural affective nuance.

Third, the current personalization design focuses primarily on static demographic and contextual parameters such as age, culture, and playback speed, but omits more dynamic, individualized factors such as user personality, real-time emotional state, or interactive feedback mechanisms. Incorporating such dimensions would further elevate the granularity and precision of emotional personalization.

These limitations, while non-trivial, delineate promising avenues for future research. Addressing them will be crucial for the development of more refined, intelligent, and emotionally adaptive systems that can fully realize the transformative potential of personalized media experiences.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33.

Bain, M.; Nagrani, A.; Brown, A.; and Zisserman, A. 2020. Condensed Movies: Story Based Retrieval with Contextual Embeddings. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Barrett, L. F.; and Russell, J. A., eds. 2014. *The Psychological Construction of Emotion*. Guilford Publications.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Busso, C.; Bulut, M.; Lee, C. M.; Kazemzadeh, A.; Mower, E.; Kim, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42: 335–359.

Cai, X.; Yuan, J.; Zheng, R.; Huang, L.; and Church, K. 2021. Speech Emotion Recognition with Multi-Task Learning. In *Proceedings of Interspeech*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Ekman, P. 1992. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3–4): 169–200.

Festinger, L. 1957. *A Theory of Cognitive Dissonance*. Row, Peterson.

Gat, I.; Aronowitz, H.; Zhu, W.; Morais, E.; and Hoory, R. 2022. Speaker Normalization for Self-supervised Speech Emotion Recognition. In *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Gu, Y.; and *et al.* 2025. Scene-Speaker Emotion Aware Network: Dual Network Strategy for Conversational Emotion Recognition. *Electronics*, 14(13): 2660.

Hofstede, G. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Sage.

Jonauskaite, D.; Abu-Akel, A.; Dael, N.; and *et al.* 2020. Universal Patterns in Color-Emotion Associations Are Further Shaped by Linguistic and Geographic Proximity. *Psychological Science*, 31(10): 1245–1260.

Li, Y.; and *et al.* 2024. Tracing Intricate Cues in Dialogue: Joint Graph Structure and Sentiment Dynamics for Multimodal Emotion Recognition. arXiv preprint arXiv:2401.01495.

Liu, C.; Chen, H.; Wang, B.; and Zheng, S. 2025. Assessing Robustness of Multi-Modal Large Language Models in Image Classification through Hierarchical WordNet-Based Evaluation. In *ICASSP*, 1–5.

Liu, J.; Liu, Z.; Wang, L.; Guo, L.; and Dang, J. 2020. Speech Emotion Recognition with Local-Global Aware Deep Representation Learning. In *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 16560–16572.

Ma, Y.; and *et al.* 2023. A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations. *IEEE Transactions on Multimedia*.

Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; and Chen, X. 2024. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 15747–15760. Bangkok, Thailand: Association for Computational Linguistics.

Marcus, A.; and Gould, E. W. 2000. Crosscurrents: Cultural Dimensions and Global Web User-Interface Design. *Interactions*, 7(4): 32–46.

Mehrabian, A. 1996. Pleasure–Arousal–Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament. *Current Psychology*, 14: 261–292.

Peng, C.; Chen, K.; Shou, L.; and Chen, G. 2024. Contrastive Feature Reconstruction and Aggregation for Multi-Modal Multi-Label Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 29374–29382.

Pussadeniya, D.; Nakisa, B.; and Rastgoo, M. N. 2024. Affective-CARA: A Knowledge Graph–Driven Framework for Culturally Adaptive Emotional Intelligence in HCI. arXiv preprint arXiv:2406.14166.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Rajapakshe, T.; Rana, R.; Khalifa, S.; Sisman, B.; Schuller, B. W.; and Busso, C. 2024. emoDARTS: Joint Optimisation of CNN & Sequential Neural Network Architectures for Superior Speech Emotion Recognition. *IEEE Access*, 12: 110492–110503.

Reeves, B.; and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.

Reiss, E. 2013. Design Dissonance: When Form and Function Collide. *UXmatters*.

Russell, J. A. 1977. Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, 11(3): 273–294.

Shen, L. 2023. Application of a Deep Generative Model for Diversified Video Subtitles Based on Generative Adversarial Networks. In *2023 3rd Asia-Pacific Conference on Communications Technology and Computer Science*.

Toshpulatov, M.; Lee, W.; Jun, J.; and Lee, S. 2025. Deep Learning Pathways for Automatic Sign Language Processing. *Pattern Recognition*, 164: 111475.

Toshpulatov, M.; Lee, W.; Lee, S.; Yoon, H.; and Kang, U. 2024. DDC3N: Doppler-Driven Convolutional 3D Network for Human Action Recognition. *IEEE Access*, 12: 93546–93567.

Valdez, P.; and Mehrabian, A. 1994. Effects of Color on Emotions. *Journal of Experimental Psychology: General*, 123(4): 394.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30.

Wu, Z.; and *et al.* 2024. Beyond Silent Letters: Amplifying LLMs in Emotion Recognition with Vocal Nuances. arXiv preprint arXiv:2407.21315.

Wu, Z.; and Palmer, M. 1994. Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*.

Yin, B.; and *et al.* 2025. Adaptive Graph Learning with Multimodal Fusion for Emotion Recognition in Conversation. *Biomimetics*, 10(7): 414.